Advancing genome annotation with long-read RNA sequencing: Insights from the IGDRion Facility

Edouard Cadieu, Aurore Besson, Matthias Lorthiois, Victor Le Bars, Armel Houel, Christophe Hitte, Catherine Andre, Thomas Derrien

Advancing genome annotation with long-read RNA sequencing: Insights from the IGDRion Facility

Edouard CADIEU[1,2], Aurore BESSON[2], Matthias LORTHIOIS[2], Victor LE BARS[1,2], Armel HOUEL[1,]
Christophe HITTE[1], Catherine ANDRE[1], Benoit HEDAN[1], Thomas DERRIEN[1,2]

1 Canine Genetics Team, CNRS, Univ Rennes – UMR6290, IGDR (Institut de Génétique et Développement de Rennes), 35043 Rennes, France

2 IGDRion, CNRS, Univ Rennes – UMR6290, IGDR (Institut de génétique et développement de Rennes), 35043 Rennes, France

Corresponding Author: thomas.derrien@univ-rennes.fr

URL: https://igdr.univ-rennes.fr/igdrion

Keywords

Long-read sequencing, Nextflow, genome annotation.

Abstract

The third revolution of sequencing, represented by the emergence of long-read technologies, enables the exploration of fundamental research questions from a new perspective [1]. In response to the growing demand for high-quality long-read sequencing data, we have established a dedicated sequencing platform called IGDRion at the IGDR institute in Rennes (https://igdr.univ-rennes.fr/igdrion). Our core facility is equipped with several long-read sequencing devices (2 MinION, 1 Mk1C, a 1 GridION and 1 P2-solo) from Oxford Nanopore Technologies (ONT) and aims to provide wet-lab and dry-lab expertise (**Fig. 1.A**).

In addition to instrumentation, the IGDRion platform utilizes established and newly developed bioinformatic programs for handling and analyzing long-read sequencing data. To adhere to FAIR principles, we employ the Nextflow-based nanoseq workflow [2] for handling Nanopore DNA/RNA sequencing data. Nanoseq is used to perform primary analyses (e.g. basecalling, demultiplexing, quality checking, alignment, …) and specific tasks (variant calling, transcript discovery, RNA quantification). In combination with nanoseq, we have also developed several bioinformatic tools which are version-controlled and freely available via a dedicated GitHub repository (https://github.com/igdrion).

In this presentation, we will demonstrate the capabilities of our long-read sequencing core facility in improving genome annotation, focusing on both coding (mRNAs) and long non-coding RNAs (lncRNAs). To achieve this, we have developed ANNEXA, an all-in-one Nextflow pipeline tailored for the analysis of LR-RNAseq sequences from ONT (**Fig. 1.B**), by incorporating the programs bambu [3] and FEELnc [4] for transcriptome discovery and lncRNAs annotation, respectively. ANNEXA facilitates the reconstruction and quantification of both known and novel genes and isoforms, while providing an extended annotation distinguishing between novel mRNA and lncRNA genes. All known and novel gene/transcript models are further characterized through structural and quantitative features (length, number of spliced transcripts, normalized expression levels...), available as graphical outputs

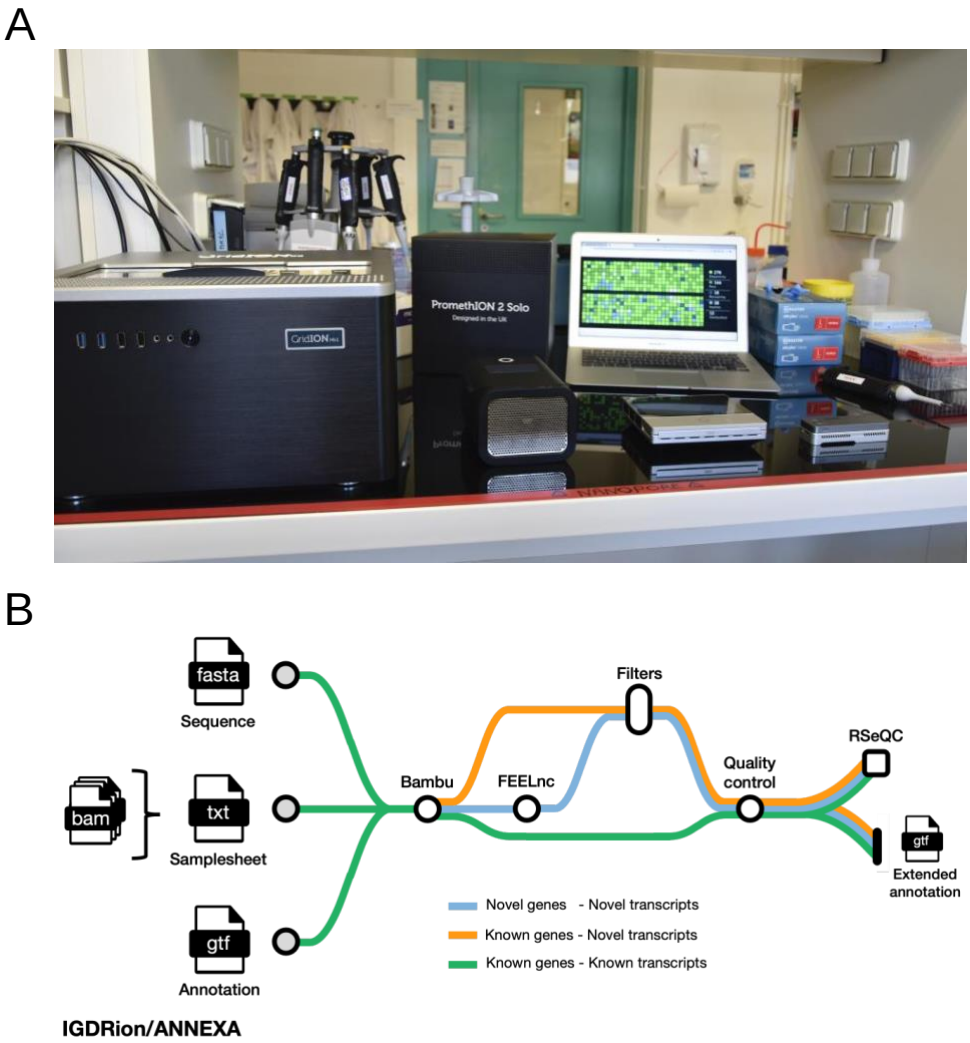

Fig. 1: ***A.*** *Sequencing instruments on the IGDRIon long-read sequencing facility (IGDR)* ***B.*** *Metro Map of the ANNEXA pipeline dedicated to the annotation of lncRNAs using long-read RNA sequencing from ONT (https://github.com/IGDRion/ANNEXA)*

.

Given the importance of assembling full-length transcripts, we have also incorporated a deep-learning method to classify and potentially filter Transcripts Start Sites (TSSs) of novel transcripts annotated by ANNEXA. Briefly, the strategy employs Encoders with self-attention mechanisms (e.g. Transformers) to classify the validity of TSS genomic sequences based on a training with species-specific reference TSS annotations.

To illustrate the utility of our programs in comparative oncology studies, we sequenced 2 human and 7 canine cancer cell lines from mucosal melanomas and histiocytic sarcomas using ONT direct cDNA sequencing, resulting in ~60 M reads (mean=6.5M reads/sample). Applying ANNEXA to these species-specific read sets, we successfully reconstructed and quantified 972 and 916 new human and canine genes, respectively, all predicted to be full-length *i.e* from their TSSs to the polyadenylation sites.

In conclusion, the IGDRion platform is a comprehensive research infrastructure integrating wet-lab and dry-lab capabilities, supported by advanced bioinformatics pipelines, to leverage the potential of long-read sequencing in advancing transcriptomic research.

Acknowledgements

Authors would like to warmly thank the Bioinformatics Genouest platform (https://www.genouest.org) for providing the required infrastructure for this work and the Ligue Régionale contre le Cancer for funding support.

References

1. Van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology A Brief History of Sequencing Technology. Trends in Genetics 2018;34:666–81.
2. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35:316–9.
3. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods 2023;20:1187–95.
4. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research 2017;45:1–12.